JAMIA Open, 2025, **8(2)**, ooaf021
https://doi.org/10.1093/jamiaopen/ooaf021
Research and Applications

Research and Applications

A proof-of-concept study for patient use of open notes with large language models

Liz Salmi, AS*[1,2], Dana M. Lewis, BA[3], Jennifer L. Clarke, MD, MPH[4], Zhiyong Dong, MS[2], Rudy Fischmann, BA[5], Emily I. McIntosh, PhD[6], Chethan R. Sarabu, MD[2,7], Catherine M. DesRoches, DrPH[2,8]

[1]Department of Women's and Children's Health, Uppsala University, 752 37 Uppsala, Sweden, [2]OpenNotes, Beth Israel Deaconess Medical Center, Boston, MA 02215, United States, [3]#OpenAPS, Seattle, WA 98101, United States, [4]Department of Neurological Surgery, University of California, San Francisco, CA 94117, United States, [5]Richmond, VA 23120, United States, [6]Guelph, ON N1E 5J5, Canada, [7]Jacobs Technion-Cornell Institute, Cornell Tech, New York, NY 10044, United States, [8]Harvard Medical School, Boston, MA 02115, United States

*Corresponding author: Liz Salmi, AS, OpenNotes, Beth Israel Deaconess Medical Center, 133 Brookline Avenue, HVMA Annex, Suite 2200, Boston, MA 02215, United States (lsalmi@bidmc.harvard.edu)

L. Salmi and D.M. Lewis contributed equally to this work.

Abstract

Objectives: The use of large language models (LLMs) is growing for both clinicians and patients. While researchers and clinicians have explored LLMs to manage patient portal messages and reduce burnout, there is less documentation about how patients use these tools to understand clinical notes and inform decision-making. This proof-of-concept study examined the reliability and accuracy of LLMs in responding to patient queries based on an open visit note.

Materials and Methods: In a cross-sectional proof-of-concept study, 3 commercially available LLMs (ChatGPT 4o, Claude 3 Opus, Gemini 1.5) were evaluated using 4 distinct prompt series—*Standard*, *Randomized*, *Persona*, and *Randomized Persona*—with multiple questions, designed by patients, in response to a single neuro-oncology progress note. LLM responses were scored by the note author (neuro-oncologist) and a patient who receives care from the note author, using an 8-criterion rubric that assessed *Accuracy*, *Relevance*, *Clarity*, *Actionability*, *Empathy/Tone*, *Completeness*, *Evidence*, and *Consistency*. Descriptive statistics were used to summarize the performance of each LLM across all prompts.

Results: Overall, the Standard and Persona-based prompt series yielded the best results across all criterion regardless of LLM. Chat-GPT 4o using Persona-based prompts scored highest in all categories. All LLMs scored low in the use of *Evidence*.

Discussion: This proof-of-concept study highlighted the potential for LLMs to assist patients in interpreting open notes. The most effective LLM responses were achieved by applying *Persona*-style prompts to a patient's question.

Conclusion: Optimizing LLMs for patient-driven queries, and patient education and counseling around the use of LLMs, have potential to enhance patient use and understanding of their health information.

Lay Summary

Large language models (LLMs) are tools powered by artificial intelligence that can write text responses based on questions. They are becoming more popular among patients and doctors. While doctors are exploring how these tools can help manage tasks like responding to patient messages, there's less research on how patients might use LLMs to understand their medical notes or make health decisions.

In this proof-of-concept study, we tested 3 different LLMs to see how well they could answer patient questions about a doctor's progress note from a neuro-oncology visit (a type of care for brain tumors). Patients created the questions, and the answers from the LLMs were graded by both the doctor who wrote the note and a patient using 8 criteria, such as accuracy, clarity, and tone.

We found that responses were best when the LLMs were given "Persona-style" prompts, meaning the tool was asked to answer as if it were a doctor.

This research shows that LLMs could help patients better understand their health information, but offering assistance to patients to use these tools wisely is key to making them helpful.

Key words: generative AI; open notes; large language models; patient portals.

Introduction

Artificial intelligence (AI) tools, including large language models (LLMs), have garnered interest as a means to advance clinical research and reduce clinician workload in healthcare settings.[1–5] Following the COVID-19 pandemic, and the enforcement of the 21st Century Cures Act Information Blocking rule, health systems with patient portals experienced a surge in patient-initiated messages and increasing clinician workload.[6–10] Clinicians raised concerns about the effect increased messaging had on physician burnout.[11] In response, some health systems began charging patients for advice provided through the portal,[12] while others worried about potential ethical ramifications of charging fees for message exchange.[13] Researchers have begun deploying and

Received: January 7, 2025; **Revised:** February 26, 2025; **Editorial Decision:** March 3, 2025; **Accepted:** March 10, 2025
© The Author(s) 2025. Published by Oxford University Press on behalf of the American Medical Informatics Association. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/jamiaopen/article/8/2/ooaf021/8109288 by BETH ISRAEL DEACONESS MEDICAL CENTER user on 09 April 2025

testing AI tools to alleviate the burden of messaging and explore potential benefits of AI for clinicians.[14] One study found that patients perceived responses from an AI chatbot as more empathetic and comprehensive than those from their clinicians.[14] A separate study found chatbot responses to diverse medical queries were largely accurate, with improvements over time.[15]

Although patient uses of health AI have not been as extensively covered in academic journals as clinician use cases, patients are also adopting AI tools to understand their health.[16] A recent KFF poll found that 17% of adults in the United States said they used a chatbot at least once a month to find information about their health,[16] and individuals with chronic conditions report using commercial LLMs to augment their "thinking and decision-making" when facing medical challenges.[17]

Patient use of their health information to gain understanding and make decisions about their care, and the care of loved ones, is not a new idea. Over the prior decade, research in the domain of "open notes" demonstrated that when patients have access to their progress notes, they felt more engaged in care,[18] were more likely to take medications as prescribed,[19] and shared notes with their care partners.[20] Patients who read their notes tend to ask more informed questions, however, notes can be confusing for some patients.[21] LLMs have the potential to help with this confusion.

While there is an increasing evidence base being developed around the ethical, legal, regulatory, and technical deployments of generative AI at the health systems level,[22–28] the effectiveness and reliability of LLMs for responding to queries based on open notes remain largely unexplored from the patient point of view. Much of the research surrounding generative AI in healthcare has focused on clinician-centric applications.[3] There are few examples of AI being evaluated specifically to assist patients in managing the cognitive burdens of understanding complex medical information or facilitating their care,[29] including care outside of hospital or clinic walls.[30]

Rather than relying solely on clinician-generated queries,[15] in this proof-of-concept study we attempted to address this gap by evaluating the reliability and accuracy of 3 commercially available chat-based LLMs by designing prompts that reflect on the lived experiences of patients with brain tumors in answering a series of patient-generated questions about a real neuro-oncology progress note (open note). The goal of this work was to assess the effectiveness and versatility of LLMs in responding to patient queries based on clinical notes, considering various prompt styles, as well as performance across models, from the perspective of both a patient and a neuro-oncology clinician. By leveraging a structured rubric to assess responses across different models and prompts, this study highlights the resulting differences in LLMs, and the role prompts may play for patients, when used in conjunction with open notes.

Methods

Study design

This cross-sectional proof-of-concept study employed a multi-model, multi-scenario approach to assess the performance of 3 LLMs in responding to patient queries about a single clinical note written by a neuro-oncologist at an academic medical center. The note used for this analysis was accessed by a patient through their MyChart-based patient portal at the University of California, San Francisco (Epic Systems Corporation) (Appendix S1). The Beth Israel Deaconess Medical Center Community on Clinical Investigations determined this study is not human subjects research, and a signed identifiable patient statement was submitted to this journal.

The proof-of-concept study involved 4 series of questions designed to explore LLM adaptability to different questioning strategies and instructional framings (Table 1). Three commercially available LLMs were each tested to compare their effectiveness, receiving identical series of prompts, to ensure comparability: ChatGPT (Open AI, GPT4o accessed via API on May 22, 2024), Claude (AnthropicAI, Claude-3-Opus-20240229 accessed via API on June 4, 2024), and Gemini (Alphabet/Google, Gemini-1.5-pro accessed via API on June 4, 2024). The prompting approach included specific personas as well as variations on the order in which questions were asked.

Evaluation criteria and process

Each prompt series was evaluated as an independent test with each model, conducted in a simulated new chat (via the API, with system messaging instructions) with the LLM, ensuring no carryover of context or knowledge from previous interactions. No other custom instructions nor settings, such as memory features, were used. The prompts were first developed as a list of topics (D.L.), evaluated by a person with lived expertise of a brain tumor (L.S.), slightly revised (D.L.), and reviewed again by 2 additional individuals with brain tumors (R.F., E.M.), then finalized (D.L.) into each series (Appendix Table S2). (Figure 1 of the Prompt Refinement Process.)

A detailed rubric was developed to score LLM responses, which included: Accuracy of Medical Information (*Accuracy*), Relevance to the Patient Query (*Relevance*), Clarity of Communication (*Clarity*), *Actionability*, Empathy, and Tone (*Empathy/Tone*), *Completeness*, Reference to Evidence or Clinical Guidelines (*Evidence*), and Internal Consistency (*Consistency*) (ie, staying on topic and not losing detail). Each criterion was rated on a 1-5 scale, with detailed descriptors for each score provided to the patient and clinician evaluator. The rubric was first developed independently, then compared with the PDQI-9 (a rubric for assessing the quality of clinical notes, see Appendix Table S3).[31]

Data collection

LLM responses were evaluated by a neuro-oncologist (J.C.) as well as a patient (L.S.) who receives care from the neuro-

Table 1. Large language model (LLM) prompt series and descriptions.

Prompt series	Description
Standard order	Patient questions were presented sequentially based on the order in which the information appeared in the clinical note (Appendix S1).
Randomized order	The same patient questions were presented in a randomized order.
Persona	Patient questions were presented sequentially based on the order in which the information appeared in the clinical note (matching the Standard Order series), but were prefixed with the Persona instruction, "*You are an expert oncologist who specializes in brain cancer*"
Randomized persona	The randomized order series, with the addition of the identical Persona instruction.

Downloaded from https://academic.oup.com/jamiaopen/article/8/2/ooaf021/8109288 by BETH ISRAEL DEACONESS MEDICAL CENTER user on 09 April 2025

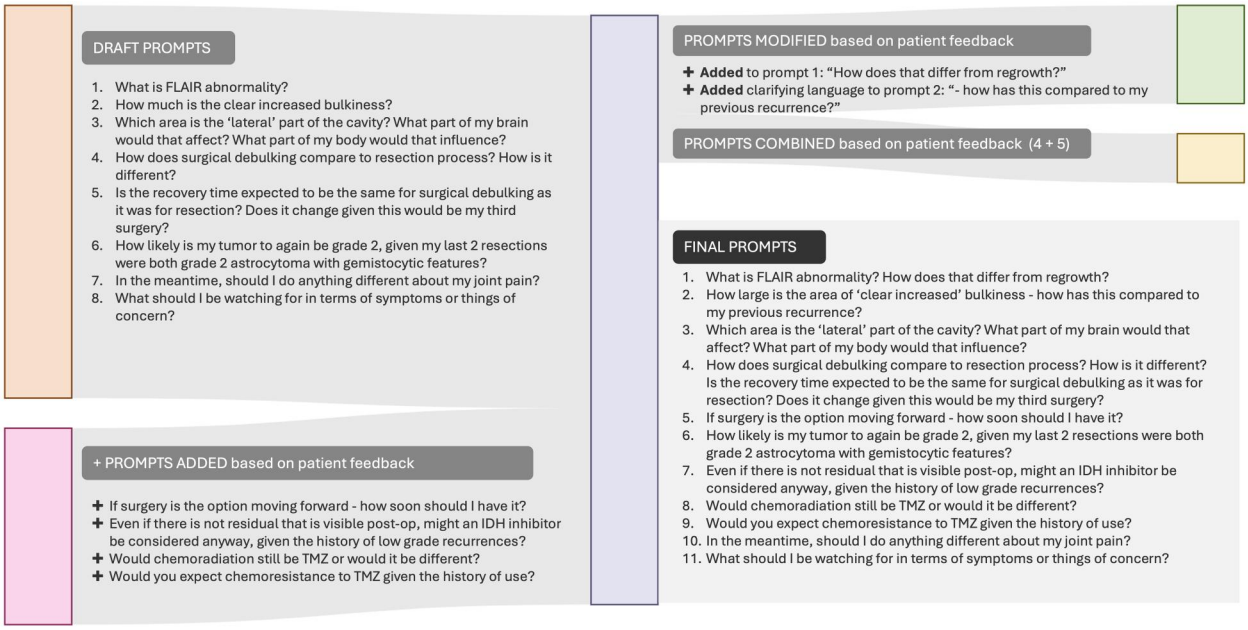

Figure 1. The prompt refinement process.

oncologist. The doctor/patient raters were blinded to which prompt series the response resulted from, as well as the LLM identity, in order to minimize bias. To further limit bias, a random order was assigned to the models and series of responses, to which the raters were blinded. The raters were encouraged to only review 3 series at a time to limit fatigue. A training session was conducted (by D.L.) to familiarize raters (J.C., L.S.) with the rubric and scoring process, which used Google Forms (Alphabet) for the blinded scoring process. Data was exported as CSV and analyzed using Python in a Jupyter Notebook.

Statistical analysis

Descriptive statistics were employed to summarize the performance of each language model across all prompts. Normality was assessed using the Shapiro-Wilk test. We used ANOVA or the Kruskal-Wallis (for non-normal data) test to identify performance differences between models and prompt series. Post-hoc analyses using Dunn's tests with Bonferroni correction for multiple comparisons were applied where applicable. Inter-rater reliability between the 2 raters (J.C., L.S.) was evaluated using Gwet's AC1 and Krippendorff's Alpha.[32] Interaction effects between model type and prompt series were analyzed using 2-way ANOVA, and interaction plots were created to visualize these effects. Finally, a heatmap was generated to visualize the average scores for each LLM response across all performance metrics.

Results

In general, the Standard Order and Persona series performed well (Figure 2 and Appendix Figure S7 and Tables S4-1 and S4-2 for score data) in this proof-of-concept study. The mean score for LLM outputs for Claude-Standard, Claude-Persona, Gemini-Standard, and ChatGPT4o-Persona was between 4.00 and 5.00 across 8 of 9 metrics—indicating a consistent performance by delivering accurate, relevant, and clear responses. Mean scores were lowest for these prompt/model combinations for the use of *Evidence*. There was a variation in *Empathy/Tone* scores across LLM outputs, with some, like Claude-Persona and ChatGPT4o-Persona, scoring higher. LLM outputs for Gemini-Standard and ChatGPT4o-Persona excelled in *Actionability* and *Completeness* (Appendices S4-1 and S4-2).

The post-hoc Dunn's test revealed significant differences in the *Accuracy* scores between Gemini and Claude ($P = .028$) and Gemini and ChatGPT4o ($P = .022$), as well as significant differences in *Relevance* between Gemini and Claude ($P = .021$). The 2-way ANOVA results showed significant interaction effects between model type and prompt series (Figure 2; $P < .001$, F-value = 6.21). Claude generally performed consistently across the prompt series, whereas Gemini had more variability. ChatGPT4o performed well in *Randomized* and *Persona* series but less so in the *Standard* series.

For models and evaluation metrics, Gemini consistently scored lower on *Relevance* across all models, whereas Claude and ChatGPT4o had higher scores. The models exhibited similar trends in *Clarity* and *Relevance*, with Claude generally leading in performance. In terms of prompt series and evaluation metrics, the *Standard* and *Persona* series generally resulted in higher scores across metrics, in contrast to the *Standard Randomized* series performing lower.

The Shapiro-Wilk test indicated that the data did not follow a normal distribution for any of the performance metrics. Kruskal-Wallis tests were used due to non-normality. Significant differences were found for *Empathy/Tone* ($P = .001$), reference to *Evidence* ($P < .001$), and internal *Consistency* ($P < .001$) between models and prompt series.

There are notable differences in the distribution of scores between the 2 raters across the same set of performance metrics, including the average overall score (Figure 3, Appendix Table S5). For most metrics, the patient tended to give slightly lower scores compared to the neuro-oncologist,

Downloaded from https://academic.oup.com/jamiaopen/article/8/2/ooaf021/8109288 by BETH ISRAEL DEACONESS MEDICAL CENTER user on 09 April 2025

Figure 2. Correlation between LLM model type and prompt series, the average score for individual metrics and overall score.

particularly in metrics like *Clarity*, *Empathy/Tone*, and *Evidence*. The pair were most similar on their rating of *Accuracy* and *Actionability* metrics but differed in their rating of *Evidence*, *Empathy*, *Completeness*, and *Relevance*. Gwet's AC1 calculation demonstrated that *Relevance* (AC1 = 0.53), *Completeness* (AC1 = 0.43), and *Clarity* (AC1 = 0.40) had the most cohesion across raters, though only *Consistency* (AC1 = 0.33, $P = .016$) showed statistically significant agreement. *Evidence* (AC1 = −0.23, $P = .006$) and *Empathy/Tone* (AC1 = 0.00, $P = .002$) exhibited the least agreement between raters, both significantly different from chance. Intraclass Correlation Coefficients (ICCs) were calculated to assess the reliability of their ratings: the patient ICC was 0.523, while the neuro-oncologist ICC was 0.618, indicating both had moderate reliability as raters. The Kruskal-Wallis test results for differences in ratings across different models and prompt series showed no statistically significant differences for any of the metrics.

Discussion

New AI tools have potential to aid patients by providing additional avenues for inquiry and understanding. We found no significant differences in the variability of ratings between models or prompt series. However, significant interaction effects were found between model type and prompt series, indicating that the performance of the models was influenced by the specific combination of these factors. Claude 3 Opus and ChatGPT 4o models performed most consistently across metrics, particularly in the *Standard Order* series (where patient questions were presented sequentially as they related to the clinical note) and *Persona* series (same as the *Standard* series, but prefixed with a Persona instruction, "*You are an expert oncologist…*"). *Standard* and *Persona* series scored high in *Accuracy*, *Relevance*, and *Clarity*. Gemini 1.5 showed more variability, with lower scores in metrics such as *Reference* and *Evidence*. Post-hoc analysis highlighted differences in *Accuracy* and *Relevance* between certain models, particularly favoring Claude and ChatGPT4o over Gemini.

Collectively, the findings from this proof-of-concept study suggest that all models demonstrated some level of capability in responding to patient queries, however, the utility varied depending on the model used, the prompt series applied, and the evaluator's perspective on a particular criterion. These results highlight the importance of considering both technical performance and human factors in the evaluation of LLMs for clinical applications, and understanding that different models may have different benefits for different types of queries.

Historically, healthcare often adopted a paternalistic stance regarding patients' access to and engagement with their medical information.[33] Clinical notes serve multiple purposes—as reminders for clinicians, as documentation for billing, as reference material for patients—so increased attention to the goal of the user matters. Studies in the domain of health services research have shown that patients often forget much of what is said in clinical visits; this misremembering worsens when a patient is receiving "bad news."[34] LLMs have the potential to bridge critical gaps in patient care by serving as an

Downloaded from https://academic.oup.com/jamiaopen/article/8/2/ooaf021/8109288 by BETH ISRAEL DEACONESS MEDICAL CENTER user on 09 April 2025

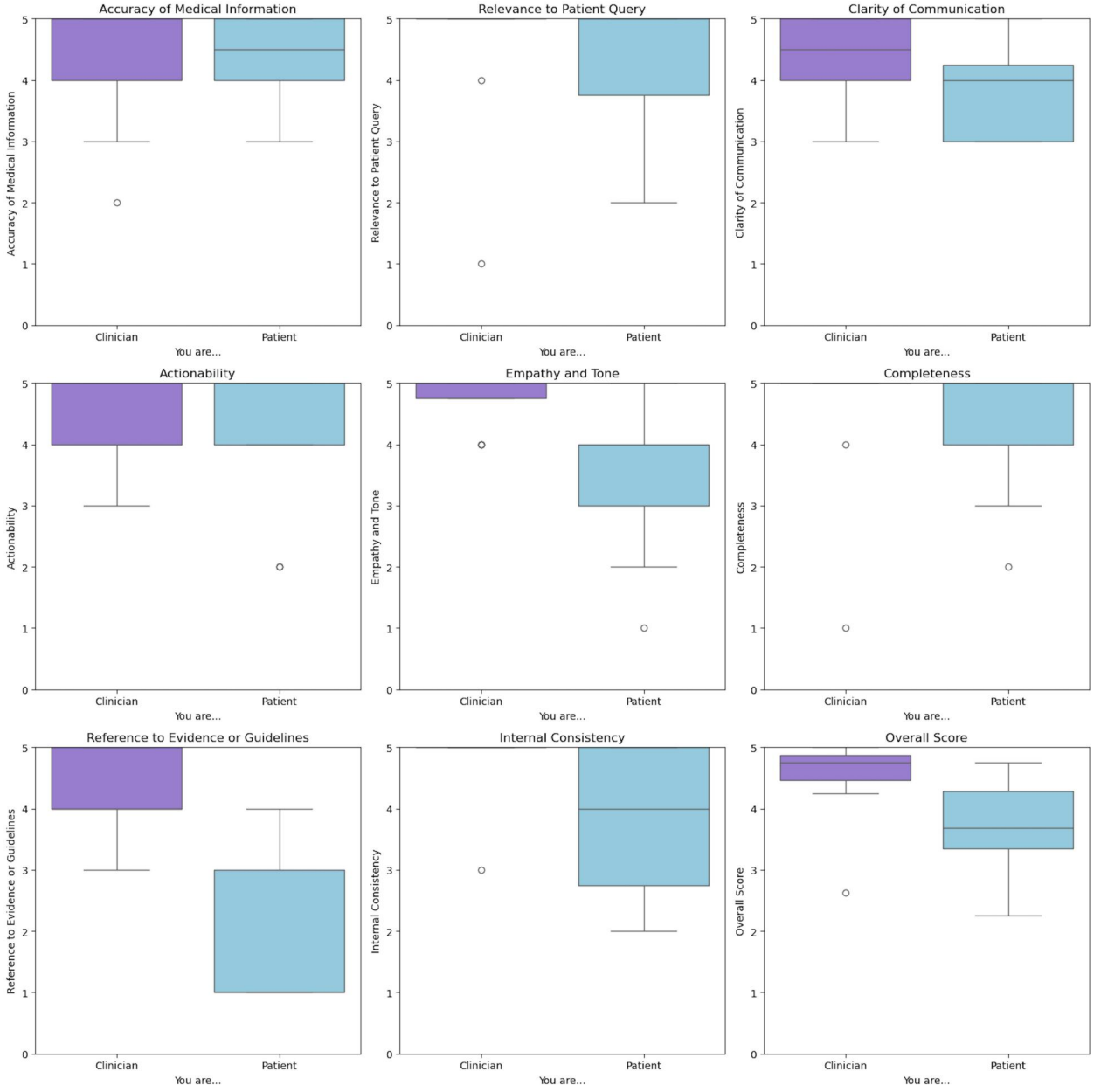

Figure 3. Average score per metric by the clinician and patient rater.

asynchronous support during "in-between" moments when patients are managing care independently.[34–36]

Previous studies have highlighted evaluations of *Empathy/Tone* and *Accuracy* when interacting with LLMs or AI tools on medical topics based on prompts designed by clinicians and researchers, however little attention was given to goals of the user (patients') interaction with the LLM.[35] For example, patient users may not prioritize the *Empathy/Tone* in an LLM response as long as they receive factual answers in response to their questions. This study showed *Accuracy* was assessed as reasonable by the patient and clinician rater across most models, whereas other users may primarily use LLMs to help assess *their own understanding* of an open note, or to help plan for a clinician follow up visit—such as demonstrated in early studies of "our notes."[36] Future LLM research should specifically evaluate criteria as it relates to the goals of the patient, especially as a growing number of consumers express concerns about AI.[27]

A proactive approach to patient education about LLMs will require a cultural shift away from discouraging patients from conducting their own search for information,[37] and rather toward helping patients better navigate the information offered by generative AI tools.[38,39] Discouraging patients from using AI may face legal challenges or accusations as being an infringement on free speech.[28] Moving forward, health systems could play an active role in patient use of AI by offering "prompting suggestions" tailored for patients who express interest in using LLMs. By doing so, healthcare providers can ensure patients are equipped with the tools necessary to engage in informed, constructive online health conversations or searches.[40–42] While clinicians should not be expected to know all details about LLMs, they should be

Downloaded from https://academic.oup.com/jamiaopen/article/8/2/ooaf021/8109288 by BETH ISRAEL DEACONESS MEDICAL CENTER user on 09 April 2025

made more aware of the range of variability among them, just as it is reasonable to expect clinicians to generally understand there are differences in medical software or medical devices. Recognizing that there are differences in LLMs, and that not all AI nor all LLMs are the same, is key.

This study did not explicitly use memory-related features—for example, the models did not retain knowledge of past queries—however this should be explored in subsequent research. Memory-related features may allow for more personalization, continuity, context-awareness, and tailored recommendations or language level-setting that would benefit patients over time.

Regulatory agencies already impose transparency requirements on generative AI as part of as part of the Assistant Secretary for Technology Policy and Office of the National Coordinator for Health Information Technology (ASTP/ONC)'s "AI assurance labs"[43,44] ASTP/ONC could also suggest collaborations with patient users of health AI in the co-development of ethical frameworks to mitigate inevitable cases of misinformation sparked through hallucinations.[45] These initiatives could guide how health systems work with, rather than against, patients in their own use of AI technologies.[46]

Limitations

Only a single clinical note was evaluated by the clinician who wrote it and the patient on whom the note is based. The clinician and patient rater have an existing clinical relationship, but they assessed the note individually. Future studies should evaluate multiple types of notes by including more evaluators without direct knowledge of the encounter covered in the note, using the same prompt series as models from this proof-of-concept study (Appendix S3). One challenge for reproducibility with LLMs is the nature of LLMs, such that the same inputs to the same models may result in slightly varying outputs. Reproduced efforts should not expect identical word output but generally similar outputs. Future studies should incorporate an assessment of the goals of the patient or user who is creating the prompts and using the LLM. For example, the prompts used in this study did not specifically prompt for evidence-based evaluation, yet this was quantified as part of the evaluation rubric, which correlates with this scoring lower compared to other metrics. Future studies should consider prompts that include evidence-based requests and/or the implication of choosing models with features enabled for search and access to an actual evidence base, rather than an LLM-only generated output.

Conclusion

In this proof-of-concept study, a neuro-oncologist and their patient evaluated the performance of 3 commercially available LLMs in responding to patient queries based on a single progress note across multiple criteria. Incorporating a Persona instruction into a prompt significantly enhanced all LLMs' performance in eliciting an empathetic yet actionable response. These findings underscore the potential of LLMs to augment patient understanding of clinical notes, but also highlight the importance of prompt design and model selection. As AI continues to evolve, future research should explore LLM performance across more diverse clinical contexts, from both patient and clinician perspectives, and the development of prompting tools to support patient users of LLMs.

Author contributions

Liz Salmi (Conceptualization, Investigation, Project administration, Visualization, Writing—original draft, Writing—review & editing), Dana M. Lewis (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing—original draft, Writing—review & editing), Jennifer L. Clarke (Data curation, Writing—review & editing), Zhiyong J. Dong (Formal analysis, Writing—review & editing), Rudy Fischmann (Data curation, Writing—review & editing), Emily I. McIntosh (Data curation, Writing—review & editing), Chethan R. Sarabu (Writing—review & editing), and Catherine M. DesRoches (Funding acquisition, Methodology, Writing—review & editing)

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This work was supported by the Patrick J. McGovern Foundation (grant #1222).

Conflicts of interest

L.S. has received speaking honoraria from Medscape. J.L.C. has received research funding from Merck and Agios/Servier, and has consulted for Agios/Servier. C.R.S. is a co-founder of OpenHand. C.M.D., L.S., and C.R.S. have received research grants from Abridge AI, Inc., and Yosemite. All other authors report no competing interests.

Data availability

The data underlying this article are available in the article and in its online Supplementary Material (Appendices S1-S7).

References

1. Tai-Seale M, Olson CW, Li J, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. *Health Aff*. 2017;36:655-662. https://doi.org/10.1377/hlthaff.2016.0811
2. Tai-Seale M, Baxter S, Millen M, et al. Association of physician burnout with perceived EHR work stress and potentially actionable factors. *J Am Med Inform Assoc*. 2023;30:1665-1672. https://doi.org/10.1093/jamia/ocad136
3. Gandhi TK, Classen D, Sinsky CA, et al. How can artificial intelligence decrease cognitive and work burden for front line practitioners? *JAMIA Open*. 2023;6:ooad079. https://doi.org/10.1093/jamiaopen/ooad079
4. Kelkar AH, Hantel A, Koranteng E, Cutler CS, Hammer MJ, Abel GA. Digital health to patient-facing artificial intelligence: ethical implications and threats to dignity for patients with cancer. *JCO Oncol Pract*. 2024;20:314-317. https://doi.org/10.1200/op.23.00412
5. Goh E, Gallo RJ, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nat Med*. Published online February 17, 2025. https://doi.org/10.1038/s41591-025-03586-x

Downloaded from https://academic.oup.com/jamiaopen/article/8/2/ooaf021/8109288 by BETH ISRAEL DEACONESS MEDICAL CENTER user on 09 April 2025

6. Salmi L, Blease C, Hägglund M, Walker J, Desroches CM. US policy requires immediate release of records to patients. *BMJ*. 2021;372:n426. https://doi.org/10.1136/bmj.n426
7. Turer RW, Desroches CM, Salmi L, Helmer T, Rosenbloom ST. Patient perceptions of receiving COVID-19 test results via an online patient portal: an open results survey. *Appl Clin Inform*. 2021;12:954-959. https://doi.org/10.1055/s-0041-1736221
8. Steitz BD, Turer RW, Lin CT, et al. Perspectives of patients about immediate access to test results through an online patient portal. *JAMA Netw Open*. 2023;6:e233572. https://doi.org/10.1001/jamanetworkopen.2023.3572
9. Hansen MA, Chen R, Hirth J, Langabeer J, Zoorob R. Impact of COVID-19 lockdown on patient-provider electronic communications. *J Telemed Telecare*. 2024;30:1285-1292. https://doi.org/10.1177/1357633X221146810
10. Neeman E, Lyon L, Sun H, et al. Future of teleoncology: trends and disparities in telehealth and secure message utilization in the COVID-19 era. *JCO Clin Cancer Inform*. 2022;6:e2100160. https://doi.org/10.1200/cci.21.00160
11. Tai-Seale M, Dillon EC, Yang Y, et al. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. *Health Aff*. 2019;38:1073-1078. https://doi.org/10.1377/hlthaff.2018.05509
12. Liu T, Zhu Z, Holmgren AJ, Ellimoottil C. National trends in billing patient portal messages as e-visit services in traditional medicare. *Health Affairs Scholar*. 2024;2:qxae040. https://doi.org/10.1093/haschl/qxae040
13. Sisk B. The harms and benefits of billing for patient portal messages. *Pediatrics*. 2023;152:e2023062188. https://doi.org/10.1542/peds.2023-062188
14. Baxter SL, Longhurst CA, Millen M, Sitapati AM, Tai-Seale M. Generative artificial intelligence responses to patient messages in the electronic health record: early lessons learned. *JAMIA Open*. 2024;7:ooae028. https://doi.org/10.1093/jamiaopen/ooae028
15. Goodman RS, Patrinely JR, Stone CA, et al. Accuracy and reliability of chatbot responses to physician questions. *JAMA Netw Open*. 2023;6:E2336483. https://doi.org/10.1001/jamanetworkopen.2023.36483
16. Marley Presiado A. KFF health misinformation tracking poll: artificial intelligence and health information. KFF. 2024. Accessed February 20, 2025. https://www.kff.org/health-information-and-trust/poll-finding/kff-health-misinformation-tracking-poll-artificial-intelligence-and-health-information
17. Goldberg C. Patient portal. *NEJM AI*. 2024;1:AIp2300189. https://doi.org/10.1056/aip2300189
18. Walker J, Leveille S, Bell S, et al. OpenNotes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. *J Med Internet Res*. 2019;21:e13876. https://doi.org/10.2196/13876
19. DesRoches CM, Bell SK, Dong Z, et al. Patients managing medications and reading their visit notes: a survey of OpenNotes participants. *Ann Intern Med*. 2019;171:69-71. Published online May 28, https://doi.org/10.7326/M18-3197
20. Chimowitz H, Gerard M, Fossa A, Bourgeois F, Bell SK. Empowering informal caregivers with health information: OpenNotes as a safety strategy. *Jt Comm J Qual Patient Saf*. 2018;44:130-136. https://doi.org/10.1016/j.jcjq.2017.09.004
21. Blease C, McMillan B, Salmi L, Davidge G, Delbanco T. Adapting to transparent medical records: international experience with "open notes. *BMJ*. 2022;379:e069861. https://doi.org/10.1136/bmj-2021-069861
22. Mello MM, Guha N. ChatGPT and physicians' malpractice risk. *JAMA Health Forum*. 2023;4:e231938. https://doi.org/10.1001/jamahealthforum.2023.1938
23. Lorenzi A, Pugliese G, Maniaci A, et al. Reliability of large language models for advanced head and neck malignancies management: a comparison between ChatGPT 4 and Gemini advanced. *Eur Arch Oto-Rhino-Laryngol*. 2024;281:5001-5006. https://doi.org/10.1007/s00405-024-08746-2
24. Feldman J, Hochman KA, Guzman BV, Goodman A, Weisstuch J, Testa P. Scaling note quality assessment across an academic medical center with AI and GPT-4. *NEJM Catal*. 2024;5:CAT.23.0283. https://doi.org/10.1056/CAT.23.0283
25. Umeton R, Kwok A, Maurya R, et al. GPT-4 in a cancer center–institute-wide deployment challenges and lessons learned. *NEJM AI*. 2024;1:AIcs2300191. https://doi.org/10.1056/AIcs2300191
26. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med Ethics*. 2021;22:122. https://doi.org/10.1186/s12910-021-00687-3
27. Levi SD, Ridgeway WE, Simon DA, et al. Utah Becomes First State To Enact AI-Centric Consumer Protection Law. Skadden. 2024. Accessed February 20, 2025. https://www.skadden.com/insights/publications/2024/04/utah-becomes-first-state
28. Blumenthal D, Goldberg C. Managing patient use of generative health AI. *NEJM AI*. 2025;2:AIpc2400927. https://doi.org/10.1056/AIpc2400927
29. Brewster RCL, Gonzalez P, Khazanchi R, et al. Performance of ChatGPT and google translate for pediatric discharge instruction translation. *Pediatrics*. 2024;154:e2023065573. https://doi.org/10.1542/peds.2023-065573
30. Kharko A, McMillan B, Hagström J, et al. Generative artificial intelligence writing open notes: a mixed methods assessment of the functionality of GPT 3.5 and GPT 4.0. *Digit Health*. 2024;10:20552076241291384. https://doi.org/10.1177/20552076241291384
31. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing electronic note quality using the physician documentation quality instrument (PDQI-9). *Appl Clin Inform*. 2012;3:164-174. https://doi.org/10.4338/ACI-2011-11-RA-0070
32. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol*. 2008;61:29-48. https://doi.org/10.1348/000711006X126600
33. Blease C, Salmi L, Rexhepi H, Hägglund M, DesRoches CM. Patients, clinicians and open notes: information blocking as a case of epistemic injustice. *J Med Ethics*. 2021;48:785-793. https://doi.org/10.1136/medethics-2021-107275
34. Laws MB, Lee Y, Taubin T, Rogers WH, Wilson IB. Factors associated with patient recall of key information in ambulatory specialty care visits: results of an innovative methodology. *PLoS One*. 2018;13:e0191940. https://doi.org/10.1371/journal.pone.0191940
35. Chen S, Kann BH, Foote MB, et al. The utility of ChatGPT for cancer treatment information. Accessed March 17, 2025. https://doi.org/10.1101/2023.03.16.23287316
36. Walker J, Leveille S, Kriegel G, et al. Patients contributing to visit notes: mixed methods evaluation of OurNotes. *J Med Internet Res*.2021;23:e29951. https://www.jmir.org/2021/11/e29951
37. Nature Medicine, eds. Will ChatGPT transform healthcare? *Nat Med*. 2023;29:505-506. https://doi.org/10.1038/s41591-023-02289-5
38. Murray E, Lo B, Pollack L, et al. The impact of health information on the internet on health care and the physician-patient relationship: national U.S. survey among 1.050 U.S. physicians. *J Med Internet Res*. 2003;5:e17. https://doi.org/10.2196/jmir.5.3.e17
39. Lu Q, Schulz PJ. Physician perspectives on internet-informed patients: systematic review. *J Med Internet Res*. 2024;26:e47620. https://doi.org/10.2196/47620
40. Collins SE, Lewis DM. Social media made easy: guiding patients to credible online health information and engagement resources. *Clin Diabetes*. 2013;31:137-141. https://doi.org/10.2337/diaclin.31.3.137
41. Katz MS, Anderson PF, Thompson MA, et al. Organizing online health content: developing hashtag collections for healthier internet-based people and communities. *JCO Clin Cancer Inf*. 2019;3:1-10. https://doi.org/10.1200/cci.18.00124
42. Hamidi N, Karmur B, Sperrazza S, et al. Guidelines for optimal utilization of social media for brain tumor stakeholders. *J Neurosurg*. 2022;136:335-342. https://doi.org/10.3171/2020.11.JNS203226

Downloaded from https://academic.oup.com/jamiaopen/article/8/2/ooaf021/8109288 by BETH ISRAEL DEACONESS MEDICAL CENTER user on 09 April 2025

43. Health Data, technology, and interoperability: certification program updates, algorithm transparency, and information sharing (HTI-1) final rule. HealthIT.gov. 2024. Accessed February 20, 2025. https://www.healthit.gov/topic/laws-regulation-and-policy/health-data-technology-and-interoperability-certification-program
44. Schumaker E. Artificial Intelligence Assurance Labs Are Coming, HHS chief AI officer says. POLITICO PRO. 2024. Accessed February 20, 2025. https://subscriber.politicopro.com/article/2024/09/artificial-intelligence-assurance-labs-are-coming-hhs-chief-ai-officer-says-00179815
45. Shah NH, Halamka JD, Saria S, et al. A nationwide network of health AI assurance laboratories. *JAMA*. 2024;331:245-249. https://doi.org/10.1001/jama.2023.26930
46. Hantel A, Clancy DD, Kehl KL, et al. A process framework for ethically deploying artificial intelligence in oncology. *J Clin Oncol*. 2022;40:3907-3911. https://doi.org/10.1200/jco.22.01113

Downloaded from https://academic.oup.com/jamiaopen/article/8/2/ooaf021/8109288 by BETH ISRAEL DEACONESS MEDICAL CENTER user on 09 April 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
JAMIA Open, 2025, 8, 1–8
https://doi.org/10.1093/jamiaopen/ooaf021
Research and Applications